Atomic Energy Control Board
P.O. Box 1046
Ottawa, Canada
K1P 5S9

Commission de contrôle de l'énergie atomique
C.P. 1046
Ottawa, Canada
K1P 5S9

INFO-0418

92-05775/

NAME ENCODING SCHEMES

(AECB Project No. 7.2.50)

by

P. Lalonde, M.E. Fair,
M. Carpenter and T. Scott

Statistics Canada

A research report prepared for
the Atomic Energy Control Board
Ottawa, Canada

June 1992

Canada

Research report

NAME ENCODING SCHEMES

A report prepared by Pierre Lalonde, Martha E. Fair, Maureen Carpenter and Tannys Scott, Occupational and Environmental Health Research Section, Canadian Centre for Health Information, Statistics Canada, under contract to the Atomic Energy Control Board.

ABSTRACT

A surname encoding scheme is used in all the computerized long-term follow-up studies carried out at Statistics Canada as this allows a more economical and efficient method of matching records. There are different phonetic codes in existence: the one currently being used at Statistics Canada, Canadian Centre For Health Information, is the New York State Identification and Intelligence System (NYSIIS) surname code developed in New York State.

A number of projects have indicated that at least 1-2% of mortality (death) links are lost because of problems with this name encoding scheme. Recommendations have been developed within this project for modifications to the system currently being used to orient it more towards Canadian surnames, particularly taking into account those where the French phonetics are causing problems. The main advantage will be to improve the performance and effectiveness of matching surnames and therefore the identification of increased cases for selection.

To evaluate whether other methods would increase the numbers of confirmed matched records economically the following possibilities were also reviewed: (i) the effectiveness of other available encoding schemes, for matching surnames; (ii) the effectiveness of the procedure of performing a second pass of linked files using and comparing different matching criteria; (iii) the advantages of using similar encoding schemes for matching forenames.

RÉSUMÉ

Dans toutes les études informatisées de suivi à long terme effectuées par Statistique Canada, on utilise un mode d'encodage des noms de famille qui permet de coupler les dossiers de façon plus économique et plus efficace. Il existe différents codes phonétiques : au Centre canadien de l'information sur la santé de Statistique Canada, on utilise actuellement le code de noms de famille NYSIIS (New York State Identification and Intelligence System) de l'État de New York.

Plusieurs études ont révélé que dans les cas de mortalité (décès), au moins 1 à 2 pour 100 des recherches sont infructueuses en raison des problèmes posés par ce mode d'encodage des noms. Des recommandations ont été élaborées dans le présent document dans le but de modifier le système actuel afin de l'axer davantage sur les noms de

famille canadiens, en tenant particulièrement compte des problèmes posés par les noms à consonance française. Le principal avantage consistera à améliorer le rendement et l'efficacité des opérations de couplage et, par conséquent, à augmenter le nombre de cas identifiés aux fins de sélection.

Pour déterminer si d'autres méthodes pourraient permettre d'augmenter de façon économique le nombre de confirmations de dossiers couplés, les points suivants ont aussi été étudiés : (i) l'efficacité des autres modes d'encodage pour le couplage des noms de famille; (ii) l'utilité du deuxième passage-machine des fichiers couplés, en se servant de différents critères de couplage et en les comparant; (iii) les avantages d'utiliser des modes d'encodage similaires pour le couplage des prénoms.

DISCLAIMER

The Atomic Energy Control Board is not responsible for the accuracy of the statements made or opinions expressed in this publication and neither the Board nor the author assumes liability with respect to any damage or loss incurred as a result of the use made of the information contained in this publication.

TABLE OF CONTENTS

		Page
Abstract		
A.	Introduction	1
	- Objectives	4
B.	Plan of the Study	5
	- Files Used	5
C.	Why A Surname Encoding Routine is Required	6
	- Selecting a Surname Encoding Procedure	7
D.	Comparing NYSIIS with Other Name Encoding Routines: Review and Evaluation	8
	1- Russell Soundex	8
	2- FONEM	9
	3- Lein, Roger Root, CIA Dictionary Method	9
	4- Utah Phonetic Transducer	10
	5- ONCA	10
E.	NYSIIS Revisited	12
	- Review	12
	- Tests and Results of NYSIIS Modifications	13
F.	Testing Other Second Pass Blocking Schemes	16
	1- First Given Name and Birth Year	18
	2- Soundex of the Surname	18
	3- Sex, Soundex of the Surname, and First Initial	18
	4- Sex, Soundex of the Surname, and Birth Year	19
	5- Results of Second Passes	19
G.	Given Name Comparison Techniques	21
H.	Other Ways of Improving the Linkage Success Rate	21
	1- Known Problem Surnames	23
	2- Alternate Source of Information	23
	3- NAMECHECK Progam	23
	4- SEXCHECK Program	24
	5- Results of Adding Alternate Records	25
I.	Recommendations and Future Follow-Up Action	27
	1- NYSIIS	27
	2- Other Blocking Schemes as a Second Pass	27
	3- Other Ways of Improving the Linkage Success Rate	28
	4- Given Name Comparison Techniques	29
	5- Choosing the 'Best' Blocking Information - Future Work	29
J.	References	30

APPENDICES

Appendix A NYSIIS
Appendix B Russell Soundex
Appendix C FONEM
Appendix D Utah Phonetic Transducer
Appendix E ONCA
Appendix F Known Problems with the NYSIIS Routine
Appendix G Corrections to NYSIIS Problems

A Introduction

Long term medical follow-up studies are those in which individuals or populations selected on the basis of whether they have been exposed to risk (e.g. radiation),received a specific preventive or therapeutic procedure,(e.g. mammography) or possess a certain characteristic, are followed to assess the outcome of the exposure, the procedure, or the effect of the characteristic e.g. occurrence of disease or death. This report addresses the problem of bringing together records for such studies, using names. In particular, various phonetic name encoding schemes are examined in detail.

The ultimate aim is to improve procedures for record linkage so that follow-up studies will be more accurate and better able to detect areas of risk. Projects with which Atomic Energy Control Board (AECB) have been directly involved are the Ontario Miners Study, the National Dose Registry Study, the Newfoundland Fluorspar Miners Study, the Tuberculosis Fluoroscopy study, the "Alive" follow-up study, the Social Insurance Number (S.I.N.) Evaluation study and the Morbidity follow-up study. A number of other studies with which AECB will also be indirectly concerned (i.e. other radiation-oriented studies involving Eldorado Nuclear workers, Atomic Energy of Canada employees and others) also use phonetic coding of names.

The ideal situation would be the assignment of a unique number at birth and all subsequent records belonging to that individual would be truly identified by this unique allocated number. Since this is rarely achieved, it becomes necessary to depend on other personal identifying information to bring exposure and medical histories together. Surname, given names, sex, and date of birth are commonly used items of identification. These data items or variables are not unique in a large population and are subject to many changes, errors and omissions in recording.

In linking two files or records within a single file a technique has to be used to reduce comparisons to a manageable number. It is necessary to partition the files into reasonably small "blocks" in order to minimize wasteful efforts in comparison. As the surname is usually present on all records, this has generally been used for the purpose of blocking. Use of the whole alphabetic sequence of the surname has been found to be less favourable than phonetic surname codes (Newcombe et al, 1959). Important aspects of a good name coding technique are the ability to place all variations of a given surname in the same code, and yet limit the size of the codes, and to do this at a reasonable cost. In practice there is a trade-off in attempting to minimize losses of linkage without excessively increasing cost. This document describes the initial steps required in making such an assessment.

There are different phonetic codes in existence; the one currently

being used at Canadian Centre for Health Information, Statistics Canada, is the New York State Identification and Intelligence System (NYSIIS) surname code developed in New York State. All computerized long-term follow-up studies carried out in the Occupational and Environmental Health Research Section of Statistics Canada use this surname encoding scheme. It has been found to allow a more economical and efficient blocking method for matching records than some other schemes.

During the course of over 60 mortality search studies conducted at Statistics Canada, it has been found that a number of links are lost because of problems with the choice of "blocking" items and the name encoding scheme used. The number of links lost ranges from about 2% for most studies to 20% where identifiers are very poor. It is therefore proposed to make modifications to the NYSIIS routine as well as the procedures currently being used to orient it more towards the Canadian situation, particularly where the phonetics of French surnames are causing problems.

No matter how good a name encoding scheme is, there will always be some links lost because the names are truly different or vary too much to be grouped by the coding scheme. Therefore, it is also proposed to explore secondary methods of finding the more difficult links/ matches by performing second pass linkages and by generating useful entries called "alternate" records, which are compared in addition to the original entries.

Objectives

The overall objective of this project is to improve the performance and effectiveness of name matching techniques in order to increase the number of matched pairs for epidemiological studies. This can be accomplished by:

a) modifying and improving the NYSIIS surname encoding scheme to take into account Canadian names and phonetics;

b) investigating the economy and effectiveness of performing second passes of linked files and by generating alternate records;

c) generating alternate records during the preprocessing phase (e.g. checking sex code and name spellings using a dictionary approach);

d) deriving methods to improve the matching and retrieval of records using forenames.

B. Plan of the Study

The basic plan was to identify and accumulate problem cases; develop and test corrections and suggestions for improvements; and implement the most effective ones as far as possible. These resulting improvements would then be incorporated in a rematch of the miners file to the mortality file to see how many more links could be found because of the improvements.

Files Used

A study cohort of Ontario miners was used to test the modifications and new techniques. This file of approximately 91,000 persons contains 75,000 miners from the files of the Workers Compensation Board (WCB) and 16,000 workers from the files of the Ministry of Labour (MOL). This cohort of 91,000 persons was first linked to the 1978-1985 mortality files and then to the 1986 file as an update (i.e. about 2 million deaths).

C. Why a Surname Encoding Routine is Required

The ability to link records when there are variations in the spelling of a name is very important, particularly when a personal unique identifier like the S.I.N. is not available.

A surname encoding procedure is thus basically required for two reasons:

1) to transform variations of a given surname and to place them in the same code, and
2) to reduce to a more manageable number, the number of pairwise comparisons when linking records together.

When linking two files or doing an internal linkage on a single file, the ideal situation is to compare every possible pair. It may be ideal, but it is wasteful and not practical as billions of pairwise comparisons would take place, making the processing cost prohibitive. A more logical approach would be to limit the number of comparisons within specified sub-groups or blocks. These blocks should have a very high likelihood of containing all the necessary records in order to bring together the true matches.

Among the data items that are present on files that identify individuals the best item to use to create these blocks is the surname as:

1) the surname is always present;

2) the surname is a very discriminating data item;
3) it is subject to very little change as a form of permanent identification; and
4) although subject to minor errors or variations, it is still more reliable than other data items present.

Selecting a Surname Encoding Procedure

There are many surname encoding procedures, but a good encoding procedure can be defined as one which:

1) places all variations of a given surname in the same code or block;
2) limits the size of the block, that is, limits the number of records assigned the same code;
3) creates codes that contain few dissimilar surnames; and
4) requires minimal processing time and cost.

When selecting a good encoding procedure, both reliability and discriminating power need to be considered. Reliability serves to keep the losses down, and discriminating power is needed to keep the cost down. The definition of "best" implies a balancing of one against the other.

D. Comparing NYSIIS with Other Name Encoding Routines: Review and Evaluation

NYSIIS is a phonetic code which groups together similar sounding names. Vowels, some letters, and certain combination of letters are recoded and then the name is compressed by eliminating silent, redundant, or duplicate consecutive letters (see Appendix A for the rules). The resulting code is totally alphabetical and often sounds a bit like the surnames it represents. Because vowels are kept, the NYSIIS blocks are much smaller and contain much less dissimilar surnames. The NYSIIS name encoding scheme has been used at Statistics Canada for about the past 10 years. It has served us well and has yielded a very high rate of matches, but it is still not perfect. Therefore it is useful to test its effectiveness against other available encoding schemes. The following additional phonetic based techniques were considered in this project:

1- The Russell Soundex coding is simple and has very few rules (see Appendix B). It is characterized by its disregard for all vowels and the retention of the first letter of a name. The best feature about Soundex is also its weakest point. Because of its disregard for vowels and grouping of other letters, this routine groups together far too many dissimilar surnames under the same code making the blocks very large and costly to process. Soundex would be more effective as a second pass blocking if it were used in conjunction with other

fields such as sex, first initial, or maybe birth year. Some results of using Soundex as a second pass can be seen in Section F.5 and Figure 1.

2- The FONEM routine (see Appendix C) used by the University of Quebec in Chicoutimi, was designed to handle French names; in particular, those common in the Saguenay region of Quebec. It does not work well on English and other types of names. The rules are governed by the idea of reproducing French sounds, and not altering the phonetic structure of names. The rules are fairly strict to avoid grouping together too many dissimilar names. The role of FONEM is not really that of a blocking routine but that of a pattern matching routine to determine agreement between two names. However, the FONEM information was useful to identify problems with French names and to suggest possible ways of handling some of these problems.

3- Three other routines where considered briefly but were quickly dismissed as inadequate: the Lein, the Roger Root techniques and the CIA dictionary method. These three methods were referred to in a Statistical Reporting Service (SRS), Department of Agriculture paper called "Selection of a Surname Coding Procedure for the SRS Record Linkage System" (Lynch and Arends, 1977). The Lein and Roger Root methods, like the Russell Soundex method, contain many dissimilar surnames

blocked into too few large blocks. The CIA dictionary had suitably small size pockets, but the codes were very incomplete because the dictionary did not code all surnames. Those not coded were assigned a miscellaneous code and grouped together. It is interesting to note that the Department of Agriculture research group selected the NYSIIS routine as its choice for a better routine.

4- The Utah Phonetic Transducer is another phonetic name encoding technique (see Appendix D). It was developed by Victor Welsey at the University of Utah and uses a set of substitution rules similar to NYSIIS to standardize names, but organizes these rules into four distinct levels. This technique should be studied in greater detail to determine its usefulness and effectiveness. At the very least it could provide ways to improve NYSIIS because it uses similar substitution type rules. (Reference "Compilation and Evaluation of Retrospective Mortality Data for Record Linkage" by DMS system Inc, under contract to NCI).

5- The ONCA method (Oxford Name Compression Algorithm) is simply a two stage transformation involving NYSIIS and the Russell Soundex code (see Appendix E). The first stage involves NYSIIS coding of the surname and the second stage involves soundexing the NYSIIS code. The Oxford group have made a few modifications to the NYSIIS rules. Although this method brings

together even more spelling variations of surnames, it obviously has the same disadvantages as the Soundex routine has of forming groups too large and combining too many dissimilar names. This method is better than the straight Soundex, but, as with Soundex, it would be better suited as a second pass if also blocked by sex and first initial or birth year.

E. NYSIIS Revisited

Review

Many projects have reported 1-2% of links lost because of problems with NYSIIS. Following our review, we consider this to be probably better than most of the other name encoding routines. A perfect encoding routine to group all spelling variation of surnames does not exist. The best one can hope for is to try to improve what is considered to be one of the best routines. At Statistics Canada, we have been using NYSIIS for the past 10 years and, although we consider it to be best, it is not perfect. We have over the years accumulated a list of surnames that the current NYSIIS routine cannot bring together (see Appendix F). In all fairness it should be said that most of these problem surnames also cause difficulties with the other name encoding routines.

Surname variations for the same person are usually caused by keying errors when letters are entered incorrectly, or are caused by transcription errors on the source documents because the handwriting was not perfect or the person was not asked to spell his/her name upon registration.

Typically when two surnames and their corresponding NYSIIS code get compared together, the result can be one of the following situations:

Type	NYSIIS	SURNAME	Examples (NYSIIS/SURNAME)
1	SAME	SAME	SNAT/SMITH vs SNAT/SMITH
2	SAME	VARIATION	SNAT/SMITH vs SNAT/SMYTH
3	DIFFERENT	VARIATION 1	SNAT/SMITH vs SNAD/SCHMIDT
4	DIFFERENT	VARIATION 2	SNAT/SMITH vs SNJT/SMJTH
5	DIFFERENT	DIFFERENT	SNAT/SMITH vs JAN/JONES

Type Explanations

1 - Ideal situation - all are identical.

2 - OK, surnames vary but NYSIIS able to group both names.

3 - Type where surnames vary slightly and hopefully NYSIIS can be improved to group them.

4 - Type where surnames vary slightly, such as 2 letters transposed or a single letter entered incorrectly. Most of these are very difficult if not impossible to solve.

5 - Two different surnames. These are inappropriate for any encoding routine to group. The only hope is via a second pass or to have alternate records on either or both files (e.g. maiden and married surnames).

Tests and Results of NYSIIS Modifications

A number of corrections were done to the present NYSIIS (see Appendix G). After many trials, experiments, and iterations, it

appeared doubtful that modifying NYSIIS would allow grouping of as many of the problem surnames as one would like. Some of the anticipated modifications were either difficult to program or conflicted with existing rules, i.e. causing bad side effects such as breaking apart good groups previously together.

It was decided to vary the approach of the NYSIIS modification in the following manner:

1) by modifying the NYSIIS routine, with all straight forward changes not causing bad side-effects, i.e. good changes.
2) by creating another NYSIIS routine (NYSIIS2 for example) which would handle the corrections causing side effects and other problem name configurations.

Some letters, depending on the origin of the name, can sound completely different, but NYSIIS can usually handle only one sound. For example; the letter 'Q' is changed to a 'G' in the current NYSIIS which is fine for Spanish names like RODRIQUES/RODRIGUES. In many other names, especially in French names the 'Q' has a 'K' or a hard 'C' sound as in names like DOMINIQUE/DOMINIC or PASQUALE/ PASCUAL. The current NYSIIS will not group them. There are several such situations where a letter is handled correctly for some names but incorrectly for others.

Because of increased immigration there is an ever growing mixture of surnames of different ethnic origin, making it increasingly

difficult to handle many of these new surnames.

The purpose of a NYSIIS2 routine would be to give problem surnames a second chance at getting the correct NYSIIS code. NYSIIS2 routine is not intended as a replacement of the current NYSIIS, but as a complement to be run in tandem with it.

The two routines, NYSIIS and NYSIIS2, would each generate their version of the NYSIIS code. If both generate the same code (and most will), nothing else has to be done. If, however, the two codes are different, a second or alternate record should be generated with the NYSIIS2 code. As long as the number of NYSIIS2 alternate records does not exceed 10-15% of the original file, this would still be cost effective.

	SURNAME	NYSIIS	NYSIIS2		SURNAME	CODE
A)	BOURQUE	BARG	BARC	———> A1)	BOURQUE	BARG
				———> A2)	BOURQUE	BARC
B)	BOURKE	BARC	BARC	———> B)	BOURKE	BARC

In the example above, records A and B will not link together because the NYSIIS is different. However, since NYSIIS2 is the same, A will link to B because of the alternate record A2.

If one hesitates to modify the NYSIIS routine for whatever reason, using NYSIIS2 may be advantageous, not just for Statistics Canada

but for any other agency using NYSIIS and contemplating improvements to it. One could accept the current NYSIIS routine as is, but create a NYSIIS2 routine to handle problem surnames which may be specific to a particular population or file (e.g. French phonetics).

Preliminary results using NYSIIS2 alternate records can be seen in Section H.5 and Figure 2. NYSIIS2 alternate records were generated during the linkage of the 90,000 miners to the 1986 Mortality Data Base. NYSIIS2 alternates were responsible for 17 (1.2%) additional links. It should be noted that these results are preliminary in the sense that several other planned modifications to the NYSIIS2 routine had not been implemented at that time.

F. Testing Other Second Pass Blocking Schemes

The objective of a second pass or subsequent passes is to find additional links which were missed by the first or any previous linkage pass. Because it is a second or subsequent pass it should execute faster than the first pass. The program logic of subsequent passes should be able to detect and ignore all pairs which have already been compared by a previous pass, because we require new pairs only.

Data items forming the pocket identifier for subsequent passes should be fairly reliable and discriminating. Table 1 summarizes

the characteristics of the second passes considered. Sex code and NYSIIS, the first pass pocket identifier, was added to the table for comparison purposes. For testing the effects of second passes, the miners file was linked against the 1978-1985 mortality file (1,829,489 records). Four second blocking schemes are considered here:

1) First given name and birth year
2) Soundex code of the surname
3) Sex, Soundex of the surname, and first initial
4) Sex, Soundex of the surname, and birth year.

TABLE 1. Testing Second Pass Blocking Schemes

Blocking Techniques	Total Records	Number of Blocks	Records per Block	Additional Links Found**
0 Sex and NYSIIS	1,829,489	62,277	29.4	
1 First given name and birth year	1,818,232*	131,804	13.8	141 (1.5%)
2 Soundex by itself	1,829,489	5,466	334.7	127 (1.4%)
3 Sex, Soundex, and first initial	1,820,976*	129,571	14.1	112 (1.2%)
4 Sex, Soundex, and birth year	1,828,774*	335,141	5.5	94 (1.0%)

* Some records were omitted because given name and/or birth year were missing.

** Percentages are based on 9400 expected deaths.

1- First Given Name and Birth Year

The pocket identifier for this pass is seven (7) characters long. It is formed from the first four characters of the first given name and the last three digits of the full birth year e.g. MICHAEL 1942 becomes MICH942. This method was reasonably inexpensive to process and found an additional 141 links (1.5%).

2- Soundex Code of the Surname

The pocket identifier for this pass is the regular Russell Soundex code of the surname (see Section D.1 and Appendix B for details). As an example, the Soundex code for ANDERSON is A536. This method was very expensive to process and found an additional 127 links (1.4%).

3- Sex, Soundex of the Surname, and First Initial

This pocket identifier is a subdivision of the Soundex pocket (method 2) into smaller pockets. The pocket identifier is 6 characters long and is made of the sex code, the Soundex code of surname and the first initial e.g. THOMAS ANDERSON becomes 1A536T.

This method was not actually volume tested as with the methods above. Because this method is a subset of method 2 above, it was easy to determine which of 127 links identified by Soundex alone also had agreement on sex code and first initial. It is obvious that this method will yield fewer links than Soundex by itself but there will be more and smaller blocks which make the processing cost more acceptable. This method would have found 112 links.

4- Sex, Soundex of the Surname, and Birth Year

Similar to method 3, this too is a subdivision of the Soundex pockets into smaller pockets. This new pocket is 8 characters long. It is made up of the sex code, the Soundex code of the surname and the last three digits of the full birth year e.g. THOMAS ANDERSON 1942 becomes 1A536942.

As with method 3 above, the results were obtained by examining the 127 links of method 2 (Soundex) to determine how many also agreed on sex and birth year. This method would have found 94 additional links.

5- Results of Second Passes

The first two methods worked fairly well finding additional links, 1.5% and 1.4% respectively (see Table 1 and Figure 1). The clear winner in this situation is the first method, first given name and birth year, not only because of its better results but also because of its smaller pockets which made processing very reasonable.

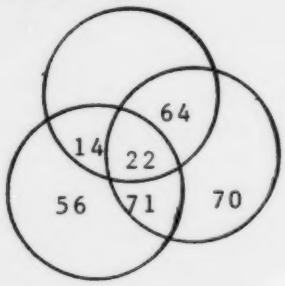
Soundex of the surname is a good second pass but the pockets are too few and too large making processing very costly. Breaking Soundex pockets into smaller pockets, as described in methods 3 and 4 above, makes it faster and more acceptable in terms of processing cost, but at the expense of lower results.

Figure. 1 Summary Results of Second Pass Linkage

Source file: 90,000 miners/workers - WCB/MOL
Target file: 1978-1985 Mortality files
Expected deaths: 9,400 approx.

Observed second pass links: 297 (3.2%) = 232 Definites
+ 65 Possibles

PASS1 (SEX + NYSIIS)
(100 links)**
(alternates)



PASS2 (GIVEN1 + BIRTH YEAR)
(227 links)

PASS3 (SOUNDEX) (163 links)

Including PASS1 alternates			Excluding PASS1 alternates		
PASS2	PASS3	PASS2+3	PASS2	PASS3	PASS2+3
Yes	No	134	Yes	No	70
No	Yes	70	No	Yes	56
Yes	Yes	93	Yes	Yes	71
227	163	297	141	127	197*
(2.4%)	(1.7%)	(3.2%)	(1.5%)	(1.4%)	(2.1%)

Reasons for failing to match during PASS1 (NYSIIS + SEX)

Sex wrong on MDB	6
" " " SIN	1
" " " WCB/MOL	1
Surname variation	249
" different	37
NYSIIS coding error	3
	297

* These 197 additional links were sent to WCB/MOL
** 100 alternate links are cases where persons had alternate records present on file. Most of these alternate records were S.I.N. Index records. Had alternate records not been present then 297 links would have been lost.

G. Given Name Comparison Techniques

Currently forenames are compared in a straightforward character by character fashion. A 'problem' given name is one where there is a nickname, short form, or spelling variations used on records relating to the same individual e.g. BILL vs. WILLIAM. To identify the most common problem given names we have assembled a file of these from 27 previous studies with over 70,000 good links. From this file of problem forenames we are creating equivalence tables or families of different spelling variations. Dr. Howard Newcombe has provided assistance in calculating appropriate weights for such tables.

The new forename comparison routine will be able to:

- find straight character agreements (as done now)
- search through equivalence tables for common spelling variations
- detect simple errors such as character inversion, skipped or wrong characters
  (e.g. ROBERT - RBOERT - RBERT - RPBERT)
- detect phonetic agreement.

H. Other Ways of Improving the Linkage Success Rate

When linking personal records, many matches are lost or are difficult to identify because of misspellings and typographical errors in people's names. Matches can also be lost if the sex code

is coded incorrectly, especially if the data files are blocked and compared separately by sex. Current methods of using name encoding techniques to group together many spelling variations of similar names work fairly well but still many spelling errors go undetected and some matches are lost.

When one thinks of ways to find more links, one usually thinks of improving or developing more sophisticated linkage techniques. Users, however, tend to forget the data, which is really the focal point of any study.

One way to find more links is to improve the quality of the data on the files that are to be linked together. When working with problem names and other problem situations to understand why links were missed, one can identify different types of errors in the data. Many of these errors could have been detected and corrected if the data had been verified. As a result of working with data containing errors, we have developed methods of improving and/or correcting the data. The data can be improved during the preprocessing phase by generating alternate records which contain more accurate information than the original information. The following types of alternate records can be generated:

1. alternates for known problem surnames;
2. alternates from a different source of information, when available (e.g. S.I.N. Index);
3. alternates with corrected name spelling (NAMECHECK);

4. alternates when sex is coded incorrectly (SEXCHECK).

1. For certain known problem surnames which cannot properly be handled by NYSIIS or other coding schemes, one can generate alternate records with the more acceptable surname spelling. A look-up table would be used to identify problem surnames.

2. When an alternate source of information can be used to complement the original file, it should be done. For example, in a cohort where the S.I.N. number is recorded, the S.I.N. Index file is an excellent source of alternate data because it contains better, more accurate identifiers. This better, more accurate data, ensures that the good links out-distance the runner-up false links by a considerably wider weight margin than they would have otherwise. It helps to an even greater degree the process of visual assessment and manual resolution. This has been tried and has proven to increase the number of good links by 2.2% (when the S.I.N. is present) in a S.I.N. evaluation study using Ontario miners. Similar results were obtained during the 1986 update linkage of the same miners file, as shown in Figure 2.

3. The NAMECHECK program is a SAS program written to detect spelling errors in name fields. It compares all surnames and given names of a study file against a master directory of surnames and given names. If the name being searched is not

found on the directory then the record is printed out with the person's names, sex, marital status and study number. A flag (->) indicates which name field is possibly in error. It is up to the user to browse through this report record by record to determine the correct or a better spelling. Once corrections are identified the user can correct the error(s) on the record or generate an alternate record with what is thought to be the correct spelling. Not all names flagged contain errors. Many are valid foreign or rare names which had not yet found their way into the master directory of names. The NAMECHECK can be responsible for generating between one and two percent of alternate records. The number of alternate records generated will depend on the quality of the data and the amount of time and effort put into resolving the problem names.

4. The SEXCHECK program is a SAS program written to detect possible errors in the sex code field. The program compares all given names in a study file against a master directory of given names. For every given name found on the directory there corresponds a score (a weight) which indicates if this given name is more oriented to a male or a female person. If the obtained total score of the person's given names suggests a sex code different from the sex code on the record, then the record is printed out with the person's names, sex, marital status, study number and the corresponding scores. It is then

up to the user to browse through the report to determine which corrections to accept. The user may then want to change the sex code, but the creation of an alternate record with the opposite sex code is recommended because, even though the score strongly suggests the opposite sex code, the original sex code could still be correct. In record linkage, it is better to enter two records when the sex code is doubtful. This is especially true if the sex code is used as part of the file blocking technique. The average number of alternate records generated because of the SEXCHECK program is about 0.5% of the source file.

5. Results of adding alternate records

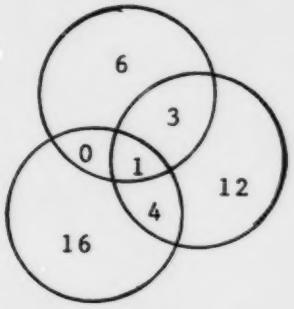
In a recent mortality study involving the linkage of 90,000 miners with the 1986 mortality file, alternate records as mentioned above were used to increase the number of links found. The expected number of deaths in 1986 for this cohort was close to 1400. This number was calculated using life expectancy tables. All results expressed in percentages are based on this number of expected deaths. Summary results are shown in Figure 2.

Figure 2. Summary Results of Adding Alternate Records

Source file: 90,000 miners/workers - WCB/MOL
Target file: 1986 Mortality file
Expected deaths: 1,400 approx.

Observed new links not involving any original WCB/MOL records
41 (2.9%) = 24 definites + 17 possibles

SIN index alternates* (10 links/.7%)



PASS2 (GIVEN1 + BIRTH YEAR)
(20 links/1.4%)

NYSIIS2 plus
Name and Sex Alternates
(21 links/1.5%)

The following table shows which type of alternate records was responsible for finding the additional links. It also explains why these links were not found using the original WCB/MOL records during the usual PASS-1 linkage which uses the regular NYSIIS and sex code to block the files.

	Type of Alternate Records					
Reasons	All	SIN*	PASS2	NYSIIS2	SURNAME	SEX
Sex wrong on MDB	1	.	1	.	.	1
" " " WCB/MOL	1	.	.	.	.	1
Surname Variation	35	8	18	17	7	.
" Different	1	.	1	.	.	.
" Same, but info. was better	3	2	.	.	1	.
TOTAL	41	10	20	17	8	2

* S.I.N. number was only available on 50% of records. The expected number of 1986 deaths for persons with a S.I.N. is approximately 500. Therefore the 10 links found represents a 2% link increase. At this rate had the S.I.N. number been available on all records, S.I.N. alternate records would have been responsible for finding 28 additional links in all.

I. Recommendations and Future Follow-up Action

1. NYSIIS

It is recommended to continue using NYSIIS as the first or main blocking factor. Because of the increasing difficulties of modifying NYSIIS, it is recommended that the modifications be done in the following manner:

a) modify NYSIIS with all straight-forward changes not causing bad side effects;

b) create another NYSIIS routine (NYSIIS2) which would handle problems or controversial corrections. NYSIIS2 is not intended as a replacement of the current NYSIIS but as a complementary routine to be run in tandem with it.

This method gives problem surnames a second chance at getting the correct NYSIIS code.

2. Other blocking schemes as a second pass

The first four characters of the first given name and the birth year proved to be a good second pass pocket identifier. It was inexpensive and found approximately 1.5% additional links. Soundex by itself as a pocket identifier found approximately 1.4% additional links but the processing was very costly, because of its large blocks. To make the use of Soundex more efficient it is

recommended, that the Soundex blocks be subdivided by also using sex code and first initial or birth year. It is also recommended not to use the direct Soundex code of a surname but to take the Soundex of the NYSIIS of the surname in order to benefit from certain letter transformations as is done with the ONCA method (Appendix E).

3. Other ways of improving the linkage success rate

Aside from modifying NYSIIS, the main blocking, and using second and multipasses, one should seek to obtain or to generate better quality data. For a cohort where the S.I.N. number is recorded, it is recommended to use the S.I.N. Index as an alternate data source to benefit from the more accurate data. In a study of Ontario miners this has been proven to increase the number of good links found by 2.2% for records where S.I.N. was present.

Other methods have been developed which can improve the quality of the data (e.g. to correct errors which could otherwise prevent potential links from forming). Sometimes links are missed because sex is coded incorrectly. To detect many of these errors, it is recommended to use the SEXCHECK program which will detect possible sex code errors. Another source of error which can be serious is the spelling of surnames and given names. To detect many errors in the spelling of names, it is recommended to use the NAMECHECK program. The SEXCHECK and NAMECHECK programs developed as a result

of this project will detect possible errors. The user can then either correct the errors or generate alternate records with what is considered to be the correct information.

4. Given name comparison techniques

Although a lot of planning and front end work has been put into improving the given name comparison techniques, not much testing has actually been done. This contract has enabled us to identify improved techniques which will improve the agreement of given names in an important way. It is therefore recommended that testing and implementation of these techniques continue outside this contract.

5. Choosing the 'best' blocking information - future work

Future work needs to be conducted to quantify the tradeoff between getting the variant forms of names into the same "block" at the price of increasing the sizes of the block, thus reducing their discriminating powers and increasing the numbers of wasted comparisons and hence the costs. Newcombe (1988) has described measures which can quantify both the favourable and the associated unfavourable consequences. These should be investigated and the relevant measures derived.

J. References

Fair M.E., Newcombe H.B., and Lalonde P. Improved Mortality Searches for Ontario Miners Using Social Insurance Index Identifiers. A research report prepared for the Atomic Energy Control Board, February 1988.

Lynch B.T., and Arends W.L. Selection of a Surname Coding Procedure for SRS Record Linkage System, U.S. Department of Agriculture, Statistical Reporting Services, Research Division, February 1977.

Lynch B.T., Arends W.L. Selection of a Surname Coding Procedure for SRS Record Linkage System, U.S. Department of Agriculture, Statistical Reporting Services, Research Division, February 1977.

Newcombe H.B., Kennedy J.M., Axford S.J. and James A.P. Automatic Linkage of Vital Records. Sci. 130: 954-959, 1959.

Newcombe H.B. Handbook of Record Linkage. Oxford University Press, 1988

APPENDIX - A

The NYSIIS Name Coding Procedure used at Statistics Canada

The NYSIIS algorithm was first developed for use by a New York State agency. The following algorithm is the algorithm as modified by Lynch and Arends (1977) of SRS, US Dept of Agriculture. An '*' indicates steps added by Lynch and Arends.

1. If the first letters of the name are:
   'MAC' then change these letters to 'MCC'
   'KN' the change the letters to 'NN'
   'K' then change these letters to 'C'
   'PH' then change these letters to 'FF'
   'PF' then change these letters to 'FF'
   'SCH' then change these letters to 'SSS'
   *'WR' then change these letters to 'RR'
   *'RH' then change these letters to 'RR'
   *'DG' then change these letters to 'GG'
   *'A,E,I,O,U' then change these letters to 'A'

*2. Drop terminal 'S' or 'Z' from all names before coding begins.

3. If the last letters of the names are:
   'EE' then change these letters to 'Y'
   'IE' then change these letters to 'Y'
   *'YE' then change these letters to 'Y'
   'DT' or 'RT' or 'RD' then change these letters to 'D'
   *'NT' or 'ND' then change these letters to 'N'
   *'IX' then change these letters to 'IC'
   *'EX' then change these letters to 'EC'
   *'JR' or 'SR' then change these letters to blank.

4. The first character of the NYSIIS code is the first character of the name after rule 1 changes.

5. In the following rules, a scan is performed on the character of the name. This is described in terms of a program loop. A pointer is used to point out the current position under consideration in the name. This step begins the loop and sets this pointer to point to the second character of the name. When all characters have been scanned go to rule 8.

6. Considering the position of the pointer, only one of the following statements can be executed.
   (a) If blank, go to rule 7. Embedded blanks are ignored.

(b) If the current position is a vowel (AEIOU) then if equal to 'EV' then change to 'AF', otherwise, change current position to 'A'.

*(c) If the current position is a Y and it is not the last letter if the name, then change the current position to an 'A'.

(d) If the current position of the letter is:
'Q' then change the letter to 'G'
'Z' then change the letter to 'S'
'M' then change the letter to 'N'

(e) If the current position is the letter 'K', then if the next letter is 'N' then replace the current position by 'N' otherwise, replace the current position by 'C'.

*(f) If the current position is the letter 'S' and the next letter 'CH' then change to 'SSA' if end of the word or change to 'SSS' if not end of word.

*(g) If the current position is the letter 'S' and the next letter 'H' then change to 'SA' if end of the word or change to 'SS' if not end of word.

(h) If the current position is the letter 'P' and the next letter 'H' then change 'PH' to 'FF'.

*(i) If the current position is the letter 'G' and the next two letters are 'HT', then change 'GHT' to 'TTT'.

*(j) If the current position is the letter 'D' and the next letter 'G' then change 'DG' to 'GG'.

*(k) If the current position is the letter 'W' and the next letter 'R' then change 'WR' to 'RR'.

(l) If the current position is the letter 'H' and either preceding or following letter is not a vowel then replace the current position with the preceding letter.

(m) If the current position is the letter 'W' and the preceding letter is a vowel then replace the current position ('W') with the preceding position.

(o) If none of these rules apply, then retain the current position letter value.

7. If the current position letter is equal to the last

letter placed in the code, then set the pointer to point to the next letter and go to step 6. The next character of the NYSIIS code is the current position letter. Increment the pointer to point at the next letter. Go to step 6.

8. If the last character of the NYSIIS code is the letter 'S', then remove it. *Not done here anymore because handled by rule 2 above.

9. If the last two characters of the NYSIIS code are the letters 'AY', then replace the letters 'AY' with the single character 'Y'.

10. If the last character of the NYSIIS code is the letter 'A', then remove this letter.

*11. If the first character of the NYSIIS code is either 'A' or space, then replace it with the first letter of the original name.

APPENDIX - B

The Russell Soundex Code

Soundex is a method which was originally designed for indexing and filing by code to group similar names regardless of spelling. The first letter of the name is retained, vowels are removed, consonants are assigned numbers from 1 to 6 to represent their sounds, and redundant code numbers are removed. The detailed rules are:

1. The first letter of the name is used in its uncoded form to serve as the prefix character of the code. (The rest of the code is numerical.)

2. Thereafter, W and H are ignored entirely.

3. A, E, I, O, U, Y are not assigned a code number, but do serve as 'separators' (see Step 5).

4. Other letters of the name are converted to a numerical equivalent:

B, P, F, V	--->	1
D, T,	--->	3
L	--->	4
M, N	--->	5
R	--->	6
All other consonants (C, G, J, K, Q, S, X, Z)	--->	2

5. There are two exceptions: (a) letters that follow prefix letters which would, if coded, have the same numerical code, are ignored in all cases unless a 'separator' (see Step 3) precedes them. (b) The second letter of any pair of consonants having the same code number is likewise ignored, i.e. unless there is a 'separator' between them in the name.

6. The final Soundex code consists of the prefix letter plus three numerical characters. Longer codes are truncated to this length, and shorter codes are extended to it by adding zeros.

Examples:

ANDERSON, ANDERSEN	--->	A536
BERGMANS, BRIGHAM	--->	B625
BIRK, BERQUE, BIRCK	--->	B620
FISHER, FISCHER	--->	F260
LAVOIE, LEVOY	--->	L100
LLWELLYN	--->	L450

APPENDIX - C

The FONEM Routine

This routine was developed by a research team at the University of Quebec in Chicoutimi. It was designed to handle French names in particular for persons from the Saguenay region. FONEM is used as part of their record linkage system that makes possible the automatic reconstruction of the families and the genealogies. Their ongoing projects range from history and anthropology to demography and genetics.

The FONEM rules which are written in French, are not listed here, please refer to the following documents:

Bouchard Gérard, Brard Patrick, Lavoie Yolande
1981 "Un code de transcription phonétique pour la reconstitution automatique des familles saguenayennes", Population, no. 6 (novembre - décembre), pp 1085-1104

Bouchard Gérard, Brard Patrick
1979 "Code de transcription phonétique: énonce de règles", Document de travail no. 33. Janvier 1979. Programme de recherches sur la société saguenayenne (PRSS). Université du Québec à Chicoutimi.

APPENDIX - D

Utah Phonetic Transducer

The following transduction rules were developed by Victor Welsey at the University of Utah (Masters Thesis). There are four levels of transduction with the rules for each succeeding level being applied to the result of the preceeding level.

Level 1

This level allows many common name variants to be grouped together while avoiding overgrouping by maintaining syllabic integrity for the most part, such that 'Anne and Anna' are allowed to match at a level with over 6 times as many distinct values as Soundex.

Note that the transduction of the vowels precedes reduction of double letters to single. This effectively reduces combinations of A, E, I, and Y to single letters, allowing matches between: Annie and Anna
Mary and Maria
Stacey and Stacy.

Also note that the removal of the leading/trailing H, and other transductions on the first letter, will allow some similar names to be grouped together at level 1 which Soundex would never group, such as: Hannah and Ana
Elma and Alma.

An example of a complete transduction group at level 1 is:
Eyring, Harring, Hearing, Herring, and Haring.

Level 1 Rules

1. Remove any embedded blanks from the name; also punctuation marks.
2. Change a C which precedes a Y, E, or I to S.
3. Change E, Y, and I, to A.
4. Strip a leading H.
5. Remove a trailing H if it is preceded by a vowel.
6. Drop the vowel from a trailing S*N, where * is a vowel.
7. Drop the A from a leading MAC.
8. Reduce sequential multiple occurrences of a letter to a single occurrence of the same letter.

Level 2 Rules

1. Remove W before R; G or K before N; and T before S.
2. *I become *A, where * is a consonant.
3. BH or BP become B.
4. CQU becomes K.
5. CH and CK become K; CHS becomes X.
6. The D is dropped from DH, DJ, and DT.
7. GK becomes K.
8. KS becomes X.
9. The H is dropped from IH, MH, TH, FH, GH, and NH.
10. The B is dropped from PB.
11. PH goes to F.
12. QU becomes K.
13. SC becomes SK, while SCH becomes SH.
14. SV becomes SW.
15. A trailing S is removed.
16. Remove all vowels following the final consonant; also Y and W.
17. US, AUGH, EIGH, AH, AO, and AW become A; AO becomes O.
18. O followed by any vowel or W is reduced to O, as are ough and uo.
19. UH and UW become U.
20. Non-leading W's before vowels are eliminated.
21. An initial O is dropped.
22. Multiple occurrences of a letter are changed to single occurrences.

Two examples of groups of names which are equivalent at level 2 are:

- Mayrick, Mereck, Merrick, Meyrick, Meyricke, Myrick
- Gainer, Gayner, Genner.

Level 3 Rules

1. Z becomes S.
2. C, G, or X become K.
3. G becomes J.
4. F, P, and V become V.
5. T becomes D.
6. N becomes M.
7. H and W are eliminated.
8. Double occurrences of letters are reduced to single occurrences.
9. At level three a bit more noise begins to appear. Examples of groups of names equivalent at level 3 are:
   - Abel, Abell, Hayball, Hepwell, Hipwell, Ivall
   - Tweddel, Tweddale, Tweddell, Tweedatle, Tyddall.

Level 4 Rules

1. Remove all A's.
2. Change U to O.
3. Reduce all double occurrences of letters to single occurrences.

Examples of names which are grouped at level 4 are:

- Blest, Blissett, Blisset, Blissitt, Felstead
- Hurford, Orford, Rufford.

APPENDIX - E

The ONCA Name Encoding Procedure

ONCA (Oxford Name Compression Algorythm). This algorythm was developed by the Unit of Clinical Epidemiology of the University of Oxford. Names undergo a two stage transformation. In the first stage the NYSIIS code of the name is generated and in the second stage, the NYSIIS code is soundexed according to the Russell Soundex rules. The people at Oxford have made a few modifications to their version of the NYSIIS routine.

Examples of Application of ONCA compared with Soundex.

Proper Name	ONCA Stage 1 (NYSIIS)	ONCA Stage 2 (Soundex of NYSIIS)	Soundex (of Name)
SMITH	SNAT	S530	S530
SMITS	SNAT	S530	S532
SCHMITZ	SNAT	S530	S253
SCHMIDT	SNAD	S530	S253
CHRISTIE	CRASTY	C623	C623
KRISTY	CRASTY	C623	K623

For full Soundex rules see Appendix B.

APPENDIX - F

Known Problems with the NYSIIS Routine

Listed below are examples of surname variations which when NYSIIS-coded come out with different NYSIIS codes. Some surnames are more common than others, while some may be quite rare. This list is by no means complete, nor is it in any meaningful order.

Many of the spelling variations shown in this appendix have never actually happened in our 10 years of matching records, many of them were visualized as potential problems. Even though the list may seem long and it contains common variations for certain names, one should not be alarmed because these types of problem variations, when they actually occur, account for only 1 or 2% of missed links in linkage applications. In all fairness to NYSIIS, it should be said that a good majority of these problem names also cause difficulties to other name encoding routines.

A) Names ending with (Y,EE,IE,YE) vs (I,E,IGH,EA)

	Surnames		NYSIIS codes	
e.g.	MUSKIE, MUSKY,			
	MUSKEE	- MUSKI, MUSKE	MASCY	- MASC
	HENRY	- HENRI	HANRY	- HANR
	LAVALLEE	- LAVALLE	LAVALY	- LAVAL
	LEE	- LEA, LEIGH	LY	- L, LAG

B) Names ending with (O,AU,EAU) vs (AULT,AULD,AUD,AUT,AUX)

e.g.	RENO, RENAU		REN
		- RENAULT	- RANALT
		- RENAULD	- RANALD
		- RENAUD	- RANAD
		- RENAUT	- RANAT
		- RENAUX	- RANAX

These are mainly names of french origin ending with the sound 'O'. Names with such ending are probably a bigger problem then others because each one of the above names falls into a different NYSIIS pocket meaning none of these variations would ever get compared with each other under the current NYSIIS rules.

C) Names with semi-silent consonants (T,P,D,L,)

e.g.	THOMPSON	- THOMSON	TANPSAN	- TANSAN
	JOHNSTON	- JOHNSON	JANSTAN	- JANSAN
	JANTZEN	- JANZEN	JANTSAN	- JANSAN
	CAMPBELL	- CAMBELL	CANPBAL	- CANBAL
	DEMPSEY	- DEMSEY	DANPSY	- DANSY
	WESTCOTT	- WESCOTT	WASTCAT	- WASCAT
	FINDLAY	- FINLAY	FANDLY	- FANLY
	CAULFIELD	- CAUFIELD	CALFALD	- CAFALD
	HOLMES	- HOMES	HALN	- HAN

D) Names with semi-silent vowels

e.g.	MISENER	- MISNER	MASANAR	- MASNAR
	FOREMAN	- FORMAN	FARAMAN	- FARMAN
	GARDINER	- GARDNER	GARDANAR	- GARDNAR
	CARLETON	- CARLTON	CARLATAN	- CARLTAN

E) The 'HAZLEWOOD - HAZELWOOD' group type surnames

e.g.	HAZLEWOOD	- HAZELWOOD	HASLAD	- HASALWAD
	CHAPPLE	- CHAPEL	CAPL	- CAPAL
	APPLETON	- APPELTON	APLATAN	- APALTAN

F) The 'SPECTRE - SPECTER' group type surnames

e.g.	SPECTRE	- SPECTER	SPACTR	- SPACTAR
	CENTRE	- CENTER	CANTR	- CANTAR

G) Names starting with 'MAK'

e.g.	MAKI	- MACKI	MAC	- MC
	MAKENZIE	- MACKENZIE	MACANSY	- MCANSY
	MAKENNEY	- MCKENNEY	MACANY	- MCANY

H) Names ending with 'X'

e.g.	LENNOX	- LENNOCKS	LANAX	- LANOC
	COX	- COCKS	CAX	- CAC
	COX	- COCKX	CAX	- CACX

I) The 'DIXON vs DICKSON' group type surnames

e.g.	DIXON	- DICKSON	DAXAN	- DACSAN
	DIXON	- DIXSON	DAXAN	- DAXSAN

J) Letter 'J' sounding like a vowel

e.g.	BOJUK	- BOYUK	BAJAC	-	BAC
	ANDREJ	- ANDRE	ANDRAJ	-	ANDR
	VAJDA	- VAYDA	VAJD	-	VAD

K) Letter 'T' vs 'D'

e.g.	PETERSON	- PEDERSON	PATARSAN	-	PADARSAN
	MITTLETON	- MIDDLETON	MATLATAN	-	MADLATAN

L) French names starting with 'DES'

e.g.	DESLORME	- DELORME	DASLARN	-	DALARN
	DESFORGES	- DEFORGE	DASFARG	-	DAFARG
	DESJARDIN	- DEJARDIN	DASJARDAN	-	DAJARDAN

M) Other French names with silent 'S'

e.g.	BOISVERT	-	BOIVERT	BASVAD	-	BAVAD
	LEVESQUE	-	LEVEQUE	LAFASG	-	LAFAG
	QUESNEL	-	QUENNELL	QASNAL	-	QANAL
	DUCHESNE	-	DUCHENE	DACASN	-	DACAN
	BELISLE	-	BELILLE	BALASL	-	BALAL

N) Letter 'Q' with a 'K' sound

e.g.	MCQUAIG	-	MCCUAIG	MCGAG	-	MCAG
	BOURQUE	-	BOURKE	BARG	-	BARC
	QUENVILLE	-	KENVILLE	QANVAL	-	CANVAL

O) Names ending with 'V' vs 'F'

e.g.	POPOV	-	POPOFF	PAPAV	-	PAPAF
	IVANOV	-	IVANOFF	IVANAV	-	IVANAF
	PETROV	-	PETROFF	PATRAV	-	PATRAF

P) Names ending with semi-silent consonants

e.g.	LAMB	-	LAM	LANB	-	LAN
	COOMBES	-	COOMES	CANB	-	CAN

Q) Names ending with (RDT,NDT) vs (RT,RD,NT,ND)

e.g.	BARNHARDT	- BARNHART	BARNARD	- BARNAD
		- BARNHARD		- BARNAD
		- BARNARD		- BARNAD
	BRANDT	- BRAND	BRAND	- BRAN
		- BRANT		- BRAN

R) Names ending with (RT,RD,NT,ND) vs (RTE,RDE,RTHE,NTE,NDE)

e.g.	FORD	- FORDE	FAD	- FARD
	BART	- BARTE	BAD	- BART
		- BARTH		- BART
		- BARTHE		- BART
	HOWARD	- HOWARTH	HAD	- HART
	LAPOINT	- LAPOINTE	LAPAN	- LAPANT
	LALOND	- LALONDE	LALAN	- LALAND

S) Exceptions to the current 'EV' NYSIIS rule

e.g.	PREVOST	-	PROVOST	PRAFAST	-	PRAVAST

T) The silent 'C' in names with 'CZ'

e.g.	OLESUK	-	OLESZCZUK	OLASAC	-	OLASCSAC
	WALZAK	-	WALCZAK	WALSAC	-	WALCSAC
	WASZAK	-	WASZCZUK	WASAC	-	WASCSAC

APPENDIX - G

Corrections to NYSIIS problems

Listed below is the same list of problems which is found in Appendix - F: Below each type of problem is a proposed correction, corrections shown are by no means final and are still subject to review. Also indicated is whether the correction should be done to the current NYSIIS or to the new NYSIIS2 routine.

A) Names ending with (Y,EE,IE,YE) vs (I,E,IGH,EA)

	Surnames		NYSIIS codes	
e.g.	MUSKIE, MUSKY,			
	MUSKEE	- MUSKI, MUSKE	MASCY	- MASC
	HENRY	- HENRI	HANRY	- HANR
	LAVALLEE	- LAVALLE	LAVALY	- LAVAL
	LEE	- LEA, LEIGH	LY	- L, LAG

Correction:

NYSIIS: None

NYSIIS2: Do not change to 'Y' if last letters are YE, EE, IE, and treat Y as a regular vowel even at end of name

Comment: Further analysis must be carried out before accepting these corrections because it affects many surnames by grouping them into larger pockets.

B) Names ending with (O,AU,EAU) vs (AULT,AULD,AUD,AUT,AUX)

e.g.	RENO, RENAU	- RENAULT	REN - RANALT
		- RENAULD	- RANALD
		- RENAUD	- RANAD
		- RENAUT	- RANAT
		- RENAUX	- RANAX

Correction:

NYSIIS: None

NYSIIS2: If name ends with AULT, AULD, AUD, AUT, AUX then drop LT, LD, D, T, X respectively

C) Names with embedded semi-silent consonants (T,P,D,L,)

e.g.	THOMPSON	- THOMSON		TANPSAN	- TANSAN
	JOHNSTON	- JOHNSON		JANSTAN	- JANSAN
	JANTZEN	- JANZEN		JANTSAN	- JANSAN
	HERTZING	- HERZING		HARTSANG	- HARSANG
	CAMPBELL	- CAMBELL		CANPBAL	- CANBAL
	DEMPSEY	- DEMSEY		DANPSY	- DANSY
	WESTCOTT	- WESCOTT		WASTCAT	- WASCAT
	FINDLAY	- FINLAY		FANDLY	- FANLY
	CAULFIELD	- CAUFIELD		CALFALD	- CAFALD
	HOLMES	- HOMES		HALN	- HAN

Correction:

NYSIIS: None

NYSIIS2: Difficult, still under review:
vowel/MP/B,P,K,S -> vowel/M/B,P,K,S
vowel/ND/L/vowel -> vowel/NL/vowel
vowel/NTZ,RTZ/vowel -> vowel/NZ,RZ/vowel

D) Names with embedded semi-silent vowels

e.g.	MISENER	- MISNER		MASANAR	- MASNAR
	FOREMAN	- FORMAN		FARAMAN	- FARMAN
	GARDINER	- GARDNER		GARDANAR	- GARDNAR
	CARLETON	- CARLTON		CARLATAN	- CARLTAN

Correction:

NYSIIS: None

NYSIIS2: Difficult, still under review

E) The 'HAZLEWOOD - HAZELWOOD' group like surnames

e.g.	HAZLEWOOD	- HAZELWOOD		HASLAD	- HASALWAD
	CHAPPLE	- CHAPEL		CAPL	- CAPAL
	APPLETON	- APPELTON		APLATAN	- APALTAN

Correction:

NYSIIS: None

NYSIIS2: Still under review

F) The 'SPECTRE - SPECTER' group like surnames

e.g.	SPECTRE	- SPECTER	SPACTR	- SPACTAR
	CENTRE	- CENTER	CANTR	- CANTAR

Correction:

NYSIIS: None

NYSIIS2: Still under review

G) Names starting with 'MAK'

e.g.	MAKI	- MACKI	MAC	- MC
	MAKENZIE	- MACKENZIE	MACANSY	- MCANSY
	MAKENNEY	- MCKENNEY	MACANY	- MCANY

Correction:

NYSIIS: Treat same as MAC
i.e. If first letters of name are MAK then change to MCC

NYSIIS2: Same as above

H) Names ending with 'X'

e.g.	LENNOX	- LENNOCKS	LANAX	- LANOC
	COX	- COCKS, COCKX	CAX	- CAC, CACX
	GROULX	- GROUX, GROU	GRALX	- GRAX, GR
	ROUSSEAUX	- ROUSSEAU	RASAX	- RAS

Correction:

NYSIIS: None

NYSIIS2: If 'X' is silent as in French names ending with AUX, EUX, OUX, OIX, ULX, AIX, UIX then drop the 'X' else change 'X' to 'C'

I) The 'DIXON vs DICKSON' group like surnames

e.g.	DIXON	- DICKSON		DAXAN	- DACSAN
	DIXON	- DIXSON		DAXAN	- DAXSAN
	AXELL	- AXCELL		AXAL	- AXCAL

Correction:

NYSIIS: None

NYSIIS2: Not complete, still under review

J) Letter 'J' sounding like a vowel

e.g.	BOJUK	- BOYUK		BAJAC	- BAC
	ANDREJ	- ANDRE		ANDRAJ	- ANDR
	VAJDA	- VAYDA		VAJD	- VAD
	BJORG	- BORG		BJARG	- BARG

Correction:

NYSIIS: None

NYSIIS2: Not complete, still under review.
- if ends with 'J' then remove 'J'
- if 'J' is followed by a consonant then change to 'A' to treat like a vowel
- if 'J' in position 2 and first letter is a consonant then change 'J' to 'A'

K) Letter 'T' vs 'D'

e.g.	PETERSON	- PEDERSON		PATARSAN	- PADARSAN
	MITTLETON	- MIDDLETON		MATLATAN	- MADLATAN

Correction:

NYSIIS: None

NYSIIS2: Still under review

L) French names starting with 'DES'

e.g.	DESLORME	- DELORME	DASLARN	-	DALARN
	DESFORGES	- DEFORGE	DASFARG	-	DAFARG
	DESJARDIN	- DEJARDIN	DASJARDAN	-	DAJARDAN

Correction:

NYSIIS: None

NYSIIS2: If 'DES' is followed by a vowel, 'S', 'Z' or 'CH' then leave as is, else change 'S' to 'A' to eliminate 'S'.

M) Other French names with silent 'S'

e.g.	BOISVERT	-	BOIVERT	BASVAD	-	BAVAD
	LEVESQUE	-	LEVEQUE	LAFASG	-	LAFAG
	QUESNEL	-	QUENNELL	QASNAL	-	QANAL
	DUCHESNE	-	DUCHENE	DACASN	-	DACAN
	BELISLE	-	BELILLE	BALASL	-	BALAL

Correction:

NYSIIS: None

NYSIIS2: Not completed yet, still under review
- if ends in 'ESNE', 'ISLE', 'ESQUE' then treat 'S' like a vowel

N) Letter 'Q' with a 'K' sound

e.g.	MCQUAIG	-	MCCUAIG	MCGAG	-	MCAG
	BOURQUE	-	BOURKE	BARG	-	BARC
	QUENVILLE	-	KENVILLE	QANVAL	-	CANVAL
	LUNDQUIST	-	LUNDQVIST	LANDGAST	-	LANDGVAST

Correction:

NYSIIS: 'QV' changes to 'QU'

NYSIIS2: Change all letter 'Q' to 'C'
'QV' changes to 'QU' then to 'CU'

O) Names ending with 'V' vs 'F'

e.g.	POPOV	-	POPOFF	PAPAV	-	PAPAF
	IVANOV	-	IVANOFF	IVANAV	-	IVANAF
	PETROV	-	PETROFF	PATRAV	-	PATRAF

Correction:

NYSIIS: None

NYSIIS2: 1) If name ends with OV of EV then change V to F
2) If EV any where but last then do not change V to F as does regular NYSIIS

P) Names ending with semi-silent consonants

e.g.	LAMB	-	LAM	LANB	-	LAN
	COOMBES	-	COOMES	CANB	-	CAN

Correction:

NYSIIS: None

NYSIIS2: Still under review

Q) Names ending with (RDT,NDT) vs (RT,RD,NT,ND)

e.g.	BARNHARDT	- BARNHART	BARNARD	- BARNAD
		- BARNHARD		- BARNAD
		- BARNARD		- BARNAD
	BRANDT	- BRAND	BRAND	- BRAN
		- BRANT		- BRAN

Correction:

NYSIIS: Change DT to D first, then change RD or RT to D

NYSIIS2: Same as above

R) Names ending with (RT,RD) vs (RTE,RDE,RTH,RTHE)

e.g.	FORD	-	FORDE	FAD	-	FARD
	BART	-	BARTE	BAD	-	BART
		-	BARTH		-	BART
		-	BARTHE		-	BART
	HOWARD	-	HOWARTH	HAD	-	HART
	LAPOINT	-	LAPOINTE	LAPAN	-	LAPANT
	LALOND	-	LALONDE	LALAN	-	LALAND

Correction:

NYSIIS: None

NYSIIS2: Remove such trailing E, H, or HE before applying RT/RD/NT/ND rule

S) Exceptions to the current 'EV' NYSIIS rule

e.g. PREVOST - PROVOST PRAFAST - PRAVAST

Correction:

NYSIIS: None

NYSIIS2: If EV anywhere but last then do not change V to F

T) The silent 'C' in names with 'CZ'

e.g. OLESUK - OLESZCZUK OLASAC - OLASCSAC

Correction:

NYSIIS: None

NYSIIS2: Change CZ to ZZ to remove effect of C